QUANTIFYING PREDICTION CONSISTENCY UNDER MODEL MULTIPLICITY IN TABULAR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning large language models (LLMs) on tabular data for classification can lead to the phenomenon of *fine-tuning multiplicity*, where equally well-performing models make conflicting predictions on the same input. Fine-tuning multiplicity can arise due to variations in the training process, e.g., seed, random weight initialization, retraining on a few additional or deleted data points. This raises critical concerns about the robustness and reliability of Tabular LLMs, particularly when deployed for high-stakes decision-making, such as finance, hiring, education, healthcare, etc. This work formalizes the unique challenge of fine-tuning multiplicity in Tabular LLMs and proposes a novel measure to quantify the robustness of individual predictions without expensive model retraining. Our measure quantifies a prediction's robustness by analyzing (sampling) the model's local behavior around the input in the embedding space. Interestingly, we show that sampling in the local neighborhood can be leveraged to provide probabilistic robustness guarantees against a broad class of equally-well-performing fine-tuned models. By leveraging Bernstein's Inequality, we show that predictions with sufficiently high robustness (as defined by our measure) will remain consistent with high probability. We also provide empirical evaluation on real-world datasets to support our theoretical results. Our work highlights the importance of addressing fine-tuning instabilities to enable trustworthy deployment of Tabular LLMs in high-stakes and safety-critical applications.

1 INTRODUCTION

Large language models (LLMs) are generating significant interest in high-stakes applications, e.g., finance, healthcare, etc., particularly in few-shot classification scenarios. Tabular data is prevalent in these sectors, making the development of Tabular LLMs (TabLLMs) an emerging research priority (van Breugel & van der Schaar, 2024). Recent studies have shown that TabLLMs perform commendably in scenarios with limited training data due to their transfer learning abilities (Hegselmann et al., 2023; Dinh et al., 2022; Yin et al., 2020; Yan et al., 2024; Wang et al., 2023). However, these models are often fine-tuned from large pre-trained models with millions or billions of parameters on small, proprietary datasets (Hu et al., 2021; Liu et al., 2022). This paucity of training data, combined with the large parameter space, introduces instability across fine-tuned variants, raising concerns about their trustworthy adoption in high-stakes applications.

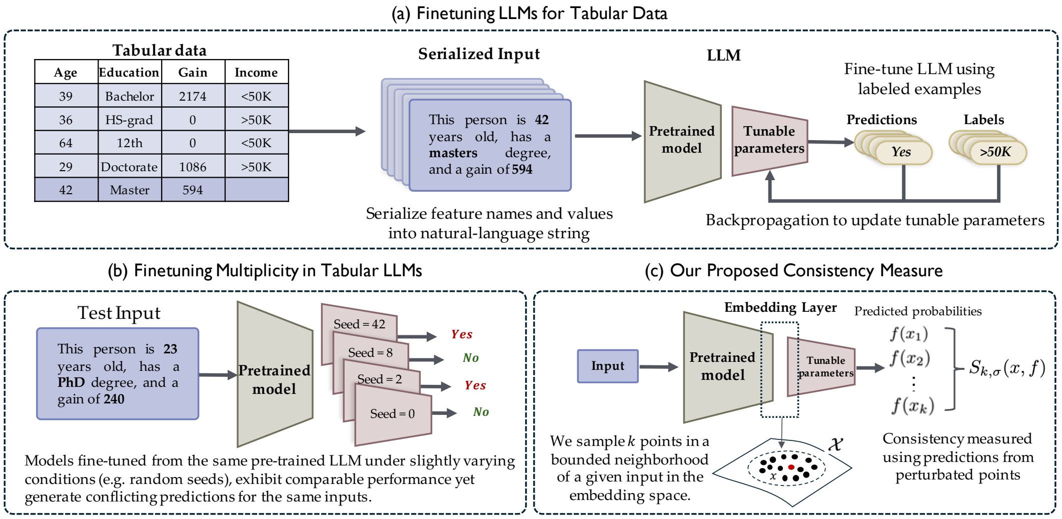
One imminent challenge is the concern of *fine-tuning multiplicity* in TabLLMs. This is the phenomenon where multiple well-performing models, fine-tuned from the same pre-trained LLM under slightly varying conditions (e.g., different random seeds or minor changes in the training data), produce conflicting predictions for the same inputs. This concept is closely related to predictive multiplicity, often referred to as the Rashomon effect in the context of neural networks (Marx et al., 2020; Breiman, 2003; Hsu & Calmon, 2022). While multiplicity has also been observed recently in LLMs in the text classification (Gomez et al., 2024), it would become particularly concerning in the context of TabLLMs for high-stakes applications. In areas like finance (Yin et al., 2023) and healthcare (Wang et al., 2024b; Chen et al., 2023b; Kim et al., 2024), arbitrary and conflicting predictions on the same input can lead to undesirable consequences, such as reputational risk and distrust.

Aside from the inherent need for predictions to be robust to minor model variations (e.g., due to different training seeds), TabLLMs deployed by institutions may also need to be updated for various reasons, e.g., to retrain on additional data points to improve performance (Wu et al., 2024), or even removing datapoints for privacy. For instance, regulatory frameworks like the GDPR (Voigt, 2017) introduce the *right to be forgotten* which necessitates the removal of an individual's data upon request, potentially leading to model updates. These updates could, in turn, impact the validity of previously issued predictions. Fine-tuning multiplicity also paves the way for fairwashing and explanation bias (Black et al., 2022; Sokol et al., 2023; Rudin et al., 2024), making quantifying robustness against fine-tuning multiplicity an important and practically relevant problem.

Existing approaches to measure multiplicity in classical machine learning often involve retraining and ensembling multiple models (Marx et al., 2020). However, such approaches can be computationally expensive for LLMs due to their large parameter sizes. This raises a key question: *Can we quantify the robustness of individual predictions without the need for expensive retraining?* To address this question, we propose a novel measure, termed *consistency*, which leverages the model's local behavior around each input data point within the embedding space to estimate the prediction's susceptibility to multiplicity. Interestingly, by analyzing this local neighborhood, we can derive probabilistic guarantees on the robustness of predictions with high consistency scores under a broad class of equally-well-performing fine-tuned models. Our contribution is summarized as follows:

- **Model multiplicity in fine-tuned Tabular LLMs.** We study the intriguing nature of fine-tuning multiplicity in Tabular LLMs. We demonstrate that prediction inconsistency exists when we actually fine-tune several models from the same pre-trained model, as observed through existing multiplicity measures such as *Arbitrariness*, *Discrepancy*, *Pairwise Disagreement*, as well as two of our proposed multiplicity measures, *Prediction Variance*, and *Range* (defined in Section 2). Furthermore, we also visualize the decision boundary for several Tabular LLMs fine-tuned for a simple classification task and unravel an interesting "noise" pattern: unlike neural network classifiers which typically have locally-smooth decision boundaries, Tabular LLMs show abrupt and impulsive variations (see Figure 2). A model having high confidence in a prediction alone does not guarantee its robustness under fine-tuning multiplicity.
- **A measure to quantify prediction robustness under fine-tuning multiplicity.** We introduce a novel measure, termed *consistency* (see Definition 5), to quantify the robustness of model predictions under fine-tuning multiplicity, without retraining several models. Given an input x and model $f(\cdot) \in (0,1)$, our robustness measure is $S_{k,\sigma}(x,f) = \frac{1}{k}\sum_{x_i \in N_{x,k}}(f(x_i) - |f(x) - f(x_i)|)$, where $N_{x,k}$ is a set of k points sampled independently from a distribution over a hypersphere of radius σ centered at x. This measure uses the input's local neighborhood (in the embedding space) to inform prediction robustness, capturing both the mean model outputs and its variability.
- **Probabilistic guarantees on consistency over a broad class of fine-tuned models.** We provide a theoretical guarantee (see Theorem 1) that predictions with sufficiently high consistency (as defined by our measure) will remain consistent with high probability over a *broad range of equally-well-performing fine-tuned models*. To achieve this guarantee, we characterize the behavior and statistical properties of this model class (see Assumption 1; Stochastic Fine-Tuned Model Class). Our results leverage Bernstein's Inequality (see Lemma 2) to derive rigorous concentration bounds used to prove our theoretical guarantee.
- **Empirical validation.** We validate our results on the Diabetes, German Credit, Bank, Heart, Car, and Adult datasets (Kahn; Hofmann, 1994; Becker & Kohavi, 1996). We employ the BigScience T0 encoder-decoder model (Sanh et al., 2021) and Google FLAN-T5 (Chung et al., 2024), fine-tuned via the T-Few recipe (Liu et al., 2022), and LORA (Hu et al., 2021). For each case, we empirically evaluate the extent of fine-tuning multiplicity, and also study how effectively our consistency measure $S_{k,\sigma}(x,f)$, (measured only using one model f) captures the multiplicity of predictions over a broad range of fine-tuned models.

Related Works: LLM in tabular predictions. The application of LLMs to tabular data is a growing area of research, demonstrating commendable performance due to the transfer learning capabilities (Yin et al., 2020; Li et al., 2020; Narayan et al., 2022; Borisov et al., 2022; Bertsimas et al., 2022; Onishi et al., 2023; Zhang et al., 2023; Wang et al., 2023; Sui et al., 2024; Yan et al., 2024; Yang et al., 2024). While neural networks and gradient boosting machines (e.g., XGBoost) perform well with tabular data when ample labeled data is available, their effectiveness drops considerably in data-scarce scenarios. In contrast, LLMs can leverage their *reasoning*, in-context learning, and pre-trained knowledge to maintain strong performance even on small, limited tabular datasets (Hegselmann et al., 2023). Dinh et al. (2022) proposes LIFT, a method for adapting LLMs to non-language



Age	Education	Gain	Income
39	Bachelor	2174	<50K
36	HS-grad	0	>50K
64	12th	0	<50K
29	Doctorate	1086	>50K
42	Master	594	

This person is **42** years old, has a **masters** degree, and a gain of **594**

This person is **23** years old, has a **PhD** degree, and a gain of **240**

Models fine-tuned from the same pre-trained LLM under slightly varying conditions (e.g. random seeds), exhibit comparable performance yet generate conflicting predictions for the same inputs.

We sample k points in a bounded neighborhood of a given input in the embedding space.

Consistency measured using predictions from perturbated points

Figure 1: (*a*) illustrates the process of fine-tuning LLMs for Tabular data using few labeled examples (Hegselmann et al., 2023; Dinh et al., 2022). (*b*) demonstrates the concept of finetuning multiplicity. Models fine-tuned from the same pre-trained LLM under slightly varying conditions, such as different random seeds, can exhibit comparable performance metrics but may yield conflicting predictions for the same input. (*c*) introduces our proposed consistency measure designed to quantify the robustness of individual predictions without requiring the retraining of multiple models. By sampling points in a bounded neighborhood around a given input in the embedding space, the consistency measure $S_{k,\sigma}(x, f)$ informs a prediction's susceptibility to multiplicity.

classification and regression tasks without changing the model architecture or loss function. Hegselmann et al. (2023) investigates the use of LLMs for zero-shot and few-shot classification of tabular data and finds that this method outperforms previous deep-learning-based approaches and is competitive with traditional baselines like gradient-boosted trees. Wang et al. (2024b) presents MediTab, a method that uses LLMs to combine different medical datasets, significantly improving predictions for patient and trial outcomes. Tabular LLMs have also been applied in other high-stakes domains (Chen et al., 2023b; Kim et al., 2024; Li et al., 2023; Yin et al., 2023). Yin et al. (2023) presents FinPT an LLM based approach to financial risk prediction. We refer to Fang et al. (2024) for a more detailed survey on LLMs for Tabular Data.

Model multiplicity in machine learning. Breiman (2003) introduced the idea that models can differ significantly while achieving similar average performance, known as the Rashomon effect. Marx et al. (2020) highlighted the prevalence of arbitrary decisions in simple classification problems, coining this phenomenon predictive multiplicity. Creel & Hellman (2022) discuss the harms of predictive multiplicity and arbitrary decisions. Methods such as TreeFarms (Xin et al., 2022), CorelsEnum (Mata et al., 2022), and RashomonGB (Hsu et al.) provide tools to enumerate models in the Rashomon set for different hypothesis spaces. Efforts to leverage model multiplicity beneficially while addressing its implications have been explored by (Black et al., 2022; Fisher et al., 2019; Xin et al., 2022; Coston et al., 2021). The effect of model multiplicity in fairness (Sokol et al., 2022) and explainability are examined by Hamman et al. (2023); Black et al. (2021); Dutta et al. (2022); Pawelczyk et al. (2020). Watson-Daniels et al. (2023); Hsu & Calmon (2022) offered a framework for measuring predictive multiplicity in classical machine learning models, however, this involves retraining several models, with the exception of Hsu et al. (2024) who propose a drop-out based approach to explore the Rashomon set for neural networks. Model multiplicity has not been extensively studied in Tabular LLMs. The closest work is by (Gomez et al., 2024), which empirically investigates prediction arbitrariness for text classification (online content moderation). In this work, we isolate and examine a specific form of multiplicity in Tabular LLMs (see Section 2). We leverage the rich embedding space of LLMs to quantify vulnerability to multiplicity without the need for expensive retraining, as fine-tuning LLMs is computationally expensive (see Section 3). There are other dimensions of robustness that focus on different aspects of model behavior such as out-of-distribution generalization, adversarial examples, and uncertainty estimation (Djolonga et al., 2020; Han et al., 2023).

1.1 Preliminaries

We consider a classification task for a tabular dataset $D = \{(x_i, y_i)\}_{i=1}^n$, where each x_i is a d-dimensional feature vector (rows of a tabular input), and each label y_i is binary, $y_i \in \{0, 1\}$. We study an n-shot classification problem by fine-tuning a pre-trained model on n examples from a training set. This fine-tuning process aims to adapt the pre-trained model to effectively predict new, unseen data points by learning from a limited number of training examples.

Serialization of Tabular Data for LLMs: To effectively apply LLMs to tabular data, it is crucial to transform the data into a natural text format. This process, known as serialization, involves converting the table rows into a text string that includes both the column names and their corresponding values (Yin et al., 2020; Jaitly et al., 2023; Hegselmann et al., 2023; Dinh et al., 2022). The resultant serialized string is combined with a task-specific prompt to form the input for the LLM. There have been various proposed methods for serialization, and this is still a topic of active research Jaitly et al. (2023). Among the serializations we have examined are: list template (a list of column names and feature values), and text template ("*The* `<column name>` *is* `<value>`."). LLMs can be adapted for classification tasks by training them on serialized tabular data. This training involves using the natural-language outputs of the LLM, mapped to valid classes in the target space, as part of a fine-tuning process (see Figure 1). To clarify, table values are serialized into serialize(x) and then transformed into a format understandable by the LLM, tokenize(serialize(x)), which is some embedding. Since these transformations are one-to-one mappings, we denote the embedded form of x as $x \in \mathcal{X}$ to represent x in the embedding space. This allows us to simplify the notation and directly use x to refer to the table values in the embedding space.

2 Model Multiplicity in Fine-Tuned Tabular LLMs

Let $f(\cdot) : \mathcal{X} \to [0, 1]$ denote an LLM that performs binary classification. We let $\mathcal{F}$ denote a broad class of competing fine-tuned models that are equally-well-performing (i.e., a set of competing models as measured by the accuracy), i.e, $\mathcal{F}_\delta = \{f : err(f) \leq err(f_0) + \delta\}$ where $err(f_0) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\hat{f}_0(x_i) \neq y_i]$ for a reference model f_0 (with satisfactory accuracy) and dataset with n examples. Here, $\hat{f}(x) = \mathbb{I}[f(x) \geq 0.5]$ denotes the predicted labels. This is a set of models that perform just as well as the baseline classifier, where $\delta \in (0, 1)$ is the error tolerance (Marx et al., 2020). The appropriate choice of δ is application-dependent.

Fine-tuning multiplicity. In this work, we explore the nature of multiplicity that arises in LLMs when fine-tuned for tabular tasks. While model multiplicity in machine learning has been studied in various contexts (see Related Works in Section 1), the unique challenges of fine-tuning multiplicity in Tabular LLMs remain relatively unexplored.

Traditional models, such as neural networks and gradient boosting machines (e.g., XGBoost), remain state-of-the-art when ample labeled data is available (Kadra et al., 2021; Gorishniy et al., 2021). However, their performance declines significantly in data-scarce scenarios. In contrast, LLMs can leverage their *reasoning* and pre-trained knowledge to achieve strong performance even with limited tabular data (few-shot learning) (Hegselmann et al., 2023). This makes LLMs appealing for few-shot tabular tasks, which often involve a mix of numerical and textual features. However, fine-tuning these large models may risk multiplicity.

To illustrate this, we conduct experiments using synthetic 2D datasets (see Figure 2). While fine-tuning an LLM on such data might seem excessive, it provides a clear visualization of the phenomenon. We fine-tune several competing models using the text template ("*The* `<column name>` *is* `<value>`") and varying only the random training seed. We reveal that fine-tuned LLMs on such non-language tasks exhibit noisy and non-smooth decision boundaries, even in regions where the model is expected to confidently predict a specifc class. We hypothesize that this noisy behavior in non-language tasks is likely because LLMs are optimized for capturing complex language structures. When fine-tuned on tabular data tasks, which often involve both text and numeric values, LLMs can leverage their pre-trained abilities but may still exhibit such instabilities.

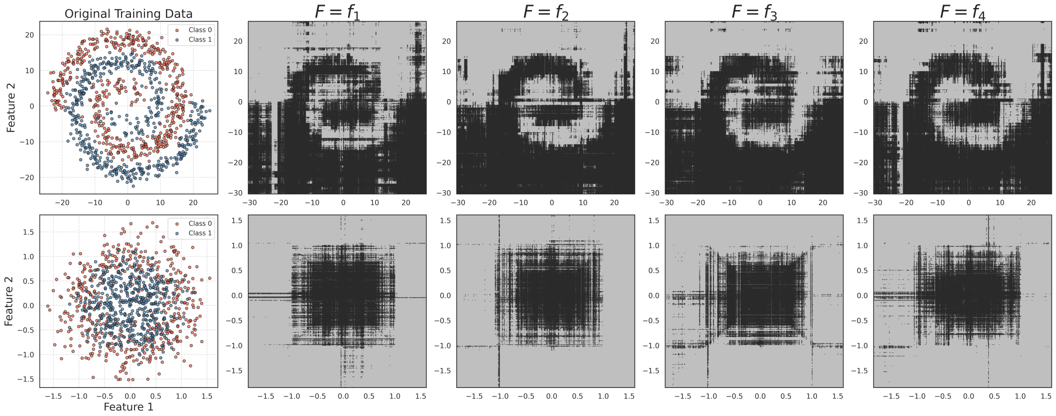


Figure 2: **Decision boundaries for multiple fine-tuned models of an LLM on synthetic datasets**. We fine-tuned several models by only changing the random training seed. All models achieve comparable training loss and accuracy, yet they converge to different functions, exhibiting intriguing noisy patterns (a phenomenon absent in models like neural networks which are typically locally-smooth). Interestingly, these noisy behaviors appear even in regions where the model is expected to confidently predict a specifc class. Observe the location and shape of these noisy patterns vary unpredictably across the various fine-tuned models, making them a possible factor contributing to prediction multiplicity. This highlights that model predictions alone may be unreliable and motivates our perturbation-based approach to quantify multiplicity.

Evaluating Fine-tuning Multiplicity. To evaluate the extent of fine-tuning multiplicity on real-world datasets, we introduce specific empirical metrics that assess how predictions may vary across different competing fine-tuned models.

Definition 1 (Arbitrariness (Gomez et al., 2024))**.** *Arbitrariness over set $\mathcal{F}_\delta$ measures the extent of conflicting predictions across the model space for a given set of inputs $\{x_1, \dots, x_n\}$. It is defined as: $A_\delta = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\exists f, f' \in \mathcal{F}_\delta, : \hat{f}(x_i) \neq \hat{f}'(x_i)]$.*

Arbitrariness generalizes the *Ambiguity* measure which computes the fraction of points where at least one model in $\mathcal{F}_\delta$ disagrees with a reference model (Marx et al., 2020). Abitrariness measures the percentage of points that receive conflicting predictions from any two models within the set $\mathcal{F}_\delta$. Arbitrariness can be defined on an input, i.e., $A(x_i) = \mathbb{I}[\exists f, f' \in \mathcal{F}_\delta, : \hat{f}(x_i) \neq \hat{f}'(x_i)]$.

Definition 2 (Discrepancy)**.** *Discrepancy quantifies the maximum proportion of conflicting predictions between the reference model and any competing model in the set. It is defined as: $D_\delta(f_0) := \max_{f \in \mathcal{F}_\delta}(\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[\hat{f}(x_i) \neq \hat{f}_0(x_i)])$.*

Discrepancy measures the maximum number of predictions that could change if a reference model is replaced with a competing model. This means that, in practice, altering multiple predictions requires that all conflicting predictions come from a single competing model.

Definition 3 (Pairwise Disagreement (Black et al., 2022))**.** *Pairwise Disagreement assesses the variability among models by measuring the proportion of instances where pairs of models within the competing set disagree: $PD_\delta(x) := \frac{1}{|\mathcal{F}_\delta|(|\mathcal{F}_\delta|-1)}\sum_{f_i, f_j \in \mathcal{F}_\delta, f_i \neq f_j} \mathbb{I}[\hat{f}_i(x) \neq \hat{f}_j(x)]$.*

Since existing measures of multiplicity focus on predicted labels, we propose more nuanced measures that leverage the predicted probabilities of model outputs:

Definition 4 (Prediction Variance)**.** *PV measures the variability of the model outputs for a given input x across different models in the set $\mathcal{F}_\delta$: $PV_\delta(x) := \frac{1}{|\mathcal{F}_\delta|}\sum_{f \in \mathcal{F}_\delta}(f(x) - \frac{1}{|\mathcal{F}_\delta|}\sum_{f' \in \mathcal{F}_\delta} f'(x))^2$.*

Prediction Variance is unaffected by accept/reject thresholds, allowing it to detect multiplicity even when predictions are consistently on one side of the decision boundary. We also define Prediction Range (PR_δ) to measure the maximum difference in model outputs for an input: $\mathrm{PR}_\delta(x) := \max_{f \in \mathcal{F}_\delta} f(x) - \min_{f \in \mathcal{F}_\delta} f(x)$. PR captures the extreme differences, providing another perspective on prediction variability.

3 A NOVEL MEASURE OF PREDICTION CONSISTENCY

Our objective is to define a measure, denoted as $S(x, f)$, for an input x and a given fine-tuned model f, that quantifies its robustness of predictions to a broad class of equally-well-performing fine-tuned models. We desire that the measure $S(x, f)$ should be high if the input x is consistent across this broad class models (see Figure 1).

Candidate Measure: Prediction probability ($S(x, f) := f(x)$). While prediction probabilities of a model $f(\cdot)$ can offer insights into its confidence in predicting a given class, they are insufficient for assessing robustness against fine-tuning multiplicity (see Table 2, Figure 3, i.e., data point with high $f(x)$ or confidence can still be susceptible to multiplicity). In our synthetic data experiments (see Figure 2), we also observe that noisy behaviors emerge in regions where the model should be confident in its predictions, leading to conflicting outcomes across various fine-tuned models. This indicates that relying solely on an input x may not provide a reliable assessment of robustness. To address this, we propose a perturbation-based approach that leverages the local neighborhood around the input x in the embedding space, ultimately leading to our theoretical measure of consistency.

3.1 PROPOSED CONSISTENCY MEASURE

Definition 5 (Consistency). *The consistency of a given prediction $f(x) \in [0, 1]$ is defined as follows:*

$$S_{k,\sigma}(x, f) := \frac{1}{k} \sum_{x_i \in N_{x,k}} f(x_i) - \frac{1}{k} \sum_{x_i \in N_{x,k}} |f(x) - f(x_i)|, \tag{1}$$

where $N_{x,k}$ is a set of k points sampled independently from a distribution over a hypersphere of radius σ centered at x, i.e., $N_{x,k} = \{x_1, x_2, \ldots, x_k\} \subset B(x, \sigma) = \{x' \in \mathcal{X} : \|x' - x\|_2 < \sigma\}$.

Remark 1. *Our consistency measure is tied to the confidence in predicting a specific class and not the predicted labels. The concept can be seamlessly applied by considering the softmax logits for predicting any given class. This also extends to multi-class classification by using the softmax logits for each class, thereby maintaining the measure's applicability across various classification tasks.*

See Appendix B for intuitions and properties of consistency measure.

3.2 THEORETICAL GUARANTEES ON CONSISTENCY

Here, we present theoretical insights that motivate and provide guarantees for our proposed robustness measure $S_{k,\sigma}(x, f)$, ensuring consistent predictions across a broad class of fine-tuned models. We represent the class of fine-tuned models by a stochastic (random) function F, such that $F \in \mathcal{F}$. We denote two random models, F and F', both of which are independently and identically distributed within $\mathcal{F}$. For clarity, we use capital letters (e.g., F, F', X_i, Z) to denote random variables, while lowercase letters (e.g., x_i, f, ϵ) indicate specific realizations. In our framework, we define a set of assumptions that delineates the behavior of a broad class of finetuned models and the statistical properties of their predictions.

Assumption 1 (Stochastic Fine-Tuned Model Class). *We assume that for any two random models $F(X)$ and $F'(X)$ are i.i.d. given an input $X = x$. Also, let the* stochastic divergence *between predictions of two random models F and F' be: $Z_i := F'(X_i) - F(X_i) - |F(X_i) - F(x)| + |F'(X_i) - F'(x)|$, where X_i is a random point sampled independently from a distribution over a hypersphere $B(x, \sigma)$. Then, we assume:* $\text{Var}\,[Z_i|F' = f', F = f] \leq \beta$ *for all $f, f' \in \mathcal{F}$.*

Intuition. The random variable Z_i captures the neighborhood stochastic divergence between predictions of two independently fine-tuned models F and F'. This captures both the difference in predictions and variability around a given point x. The first assumption ensures that F and F' provide an unbiased estimate of the prediction for x. The assumption on the variance of Z_i indicates that the variance of the stochastic neighborhood divergence within a σ-Ball of a data point between any two models' predictions is controlled. The parameter β essentially captures the similarity of the models within the local neighborhood of a data point. This concept is also somewhat analogous to the Lipschitz constant of a general function, which bounds how much the function's output can change relative to changes in its input. However, in this context, the β-bound reflects an average behavior of the models' predictions within the local neighborhood. It does not strictly enforce a uniform Lipschitz constant, especially considering that transformer models are not typically Lipschitz continuous (Kim et al., 2021) (also observe noisy non-smooth behavior in Figure 2).

Theorem 1 (Probabilistic Guarantee on Consistency)**.** *Given a data point x, a random model F' and consistency measure $S_{k,\sigma}(x, F')$. Then under Assumption 1, and $|\mathbb{E}\left[Z_i|F' = f', F = f\right]| \leq \epsilon'$, a prediction over a broad class of fine-tuned models satisfies:*

$$\Pr\left(F(x) \geq S_{k,\sigma}(x, F') - \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8(\beta + \frac{2}{3}\epsilon)}\right), \tag{2}$$

for all $\epsilon > 2\epsilon'$, The probability is over the stochastic models F and F', and the random perturbations X_i's are random points sampled independently from a distribution over a hypersphere $B(x, \sigma)$.

Theoretical guarantee on consistency interpretation. Our consistency measure $S(x, F')$ provides a probabilistic guarantee that if a data point x has a sufficiently high consistency score with respect to a random model F', then the prediction of another random model F from the same broad class of fine-tuned models will be at least $S(x, F') - \epsilon$ with high probability. For example, if $S(x, F') = 0.8$, we can be confident that $F(x)$ will be at least $0.8 - \epsilon$ with *high* probability (i.e, the prediction will remain on the positive predicted side). This implies that high consistency scores are indicative of robust predictions across different fine-tuned models. Conversely, a low consistency score does not provide significant information about the prediction's behavior, as it does not guarantee a lower bound on the prediction. For $F(x) \geq S(x, F') - \epsilon$ to hold with high probability, a large k is needed, ideally $k \gg \beta$. This implies that when β is large then more samples are needed.

Goodness of model class. The term ϵ' in our guarantee captures the quality or goodness of the fine-tuned model class. A small ϵ' indicates a well-behaved model class, suggesting that different fine-tuned models produce similar outputs in expectation within the local neighborhood of x even if predictions might vary for a given data point. Similar behavior is visualized in Figure 2, where, despite the presence of noisy variations in the decision boundaries, the local predictions around a given point remain relatively consistent across models. This behavior is expected since these models are derived from the same pre-trained model and trained with the goal of achieving similar accuracy on the dataset. In this case, *our consistency measure provides a useful and informative lower bound on the predictions $F(x)$ with a certifiably small gap*. This aligns with related formalizations, which show the existence of simpler functions within a Rashomon set, where $\sum_{x_i \in D} |f(x_i) - f'(x_i)| \leq \Delta$, across a dataset D (Semenova et al., 2022). Recent mathematical analyses of LORA also corroborate with our assumptions, such as $\mathbb{E}_X \|f(X) - f'(X)\| \leq \Delta$, for a random variable X over a bounded set (Zeng & Lee, 2023).

Conversely, a large ϵ' indicates a more erratic model class. In this case, our bound becomes less informative, and the consistency measure might perform poorly for a given point. We interpret our results as follows: The model class is not well-behaved; thus, one cannot certify a small gap between $F(x)$ and our proposed measure. We do not provide guarantees for all types of model changes, as this would be challenging with only a single model. For example, if fine-tuned models do not achieve sufficient accuracy, encounter significant variations in hyperparameter choices, or large changes in the training data, ϵ' is likely to be large. Our focus is on the multiplicity that arises due to randomness in training, such as changes in the training seed or minor adjustments in training settings (what we term the broad class of equally-well-performing fine-tuned models). In our evaluations (see Section 4), we do not assume any specific values for ϵ' and consider regular fine-tuned models without imposing any theoretical constraint. The complete proof of Theorem 1 is provided in Appendix C. Here, we include a proof sketch.

Proof Sketch: From Assumption 1, F and F' are identically distributed given X_i, hence $\mathbb{E}[F'(X_i)|X_i] = \mathbb{E}[F(X_i)|X_i]$ and $\mathbb{E}[|F'(X_i) - F'(x)||X_i] = \mathbb{E}[|F(X_i) - F(x)||X_i]$. The terms in $\mathbb{E}\left[Z\right]$ cancel each other out, resulting in $\mathbb{E}[Z] = 0$. The next step of the proof leverages the Bernstein's inequality (see Lemma 2) to provide a bound on the stochastic neighborhood divergence (see Lemma 1). The final steps of the proof leverages the reverse triangle inequality so show: $F(x) \geq \frac{1}{k}\sum_{i=1}^{k}(F(X_i) - |F(X_i) - F(x)|)$. Combining that along with Lemma 1 derives our consistency measure and guarantees. □

Table 1: Evaluated Multiplicity for Different Datasets and Number of Shots on BigScience T0. Evaluated on 40 fine-tuned models on T-Few recipe using different random seeds. Multiplicity observed in predictions across different fine-tuned model, even when models exhibit similar accuracy (in this setting $\delta = 0.02$). Fine-tuning using LORA achieves results in the same ballpark (see LORA Table 3 in Appendix D)

Dataset	No. Shots	Multiplicity Evaluation Metrics (BigScience T0)					
		Arbitrariness	Discrepancy	Avg. Pairwise Disagreement	Avg. Pred. Variance	Avg. Pred. Range	Avg. Model Accuracy
Adult	64	10%	9%	7%	0.01	0.10	83%
	128	10%	7%	8%	0.01	0.10	84%
	512	11%	8%	7%	0.01	0.12	85%
German	64	18%	10%	6%	0.01	0.20	71%
	128	17%	11%	6%	0.01	0.16	71%
	512	23%	12%	7%	0.02	0.23	72%
Diabetes	64	29%	18%	10%	0.04	0.31	71%
	128	13%	17%	11%	0.03	0.13	72%
	512	16%	16%	10%	0.02	0.18	78%
Bank	64	11%	9%	6%	0.01	0.31	66%
	128	15%	8%	7%	0.03	0.22	75%
	512	14%	8%	7%	0.02	0.16	81%
Heart	64	6%	4%	2%	0.01	0.05	78%
	128	9%	4%	3%	0.01	0.10	83%
	512	18%	7%	5%	0.01	0.19	82%
Car	64	19%	10%	6%	0.01	0.18	81%
	128	16%	7%	5%	0.01	0.14	86%
	512	8%	4%	2%	0.01	0.09	94%

4 Empirical Validation

In this section, we experiment across different datasets to *(i)* quantify the prevalence of fine-tuning multiplicity in Tabular LLMs, and *(ii)* validate the effectiveness of our proposed measure in quantifying the consistency of predictions over a broad range of equally-well-performing fine-tuned models.

Datasets and Serialization. Our experiments utilize the Diabetes (Kahn), German Credit (Hofmann, 1994), Bank (Moro et al., 2014), Heart, Car, and Adult datasets (Becker & Kohavi, 1996), serialized using the Text Template method where each tabular entry is converted into a natural language format by stating "*The* `<column name>` *is* `<value>`" This approach helps align the inputs with the training distribution of LLMs, enhancing their performance in both zero-shot and few-shot scenarios (Hegselmann et al., 2023; Dinh et al., 2022).

Models and Fine-tuning Methods. We use the BigScience T0 (Sanh et al., 2021) and Google FLAN-T5 (Chung et al., 2024) encoder-decoder models as our pretrained LLMs. T0 is specifically pre-trained for zero-shot generalization through multitask learning. FLAN-T5 is instruction fine-tuned on a diverse range of tasks, achieving strong performance in few-shot settings. These characteristics make both models well-suited for our experiments. For fine-tuning, we adopt the T-Few recipe (Liu et al., 2022), known for its effectiveness in few-shot learning, and LORA (Hu et al., 2021), a parameter-efficient method that constrains weight matrix updates to be low-rank. Detailed setup can be found in Appendix D.3.

Evaluating Extent of Fine-tuning Multiplicity. We measure the extent of fine-tuning multiplicity across the various datasets and fine-tuning methods, we use the multiplicity evaluation metrics (introduced in Section 2) To evaluate these multiplicity metrics across our datasets, we fine-tune 40 models on Tfew recipe and LORA using different random seeds and test on a sample set. Here are the experiments we conducted:

- We evaluate multiplicity on the *BigScience T0* model fine-tuned using *T-Few* (see Table 1).
- We evaluate multiplicity on *BigScience T0* fine-tuned using *LORA* (see Table 3 in Appendix D).
- We evaluate multiplicity on *Flan-T5* model fine-tuned using *T-Few* (see Table 4 in Appendix D).

Comparing Consistency Measure to Evaluated Multiplicity. We assess the utility of our proposed consistency measure $S_{k,\sigma}(x, f)$ in informing the presence of fine-tuning multiplicity. This utility is measured using the Spearman correlation coefficient (see Definition 6), between our consistency $S_{k,\sigma}(x, f)$ (estimated on just one model) and the evaluated multiplicity (evaluated on several fine-tuned models), e.g., Spearman$(S_{k,\sigma}(x, f), PV_{\delta}(x))$ across the test set.

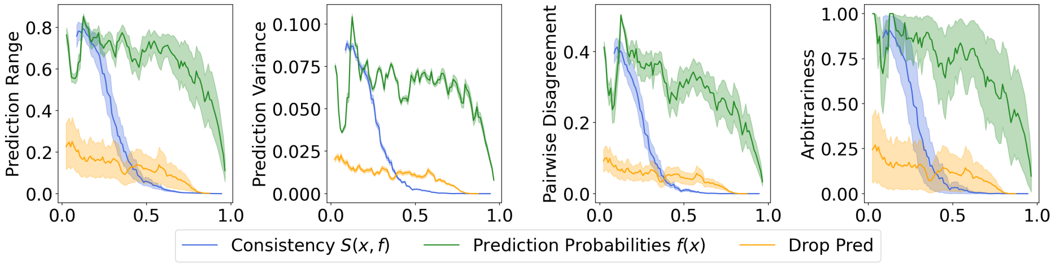


Figure 3: Evaluated multiplicity (assessed on 40 retrained models) versus our consistency measure, predicted probabilities, and drop-out method (evaluated on one model) for the 128-shot setting on the Adult dataset. The plots demonstrate that high consistency values correspond to low multiplicity across various multiplicity evaluation metrics. Also, observe that high predicted probability values (i.e., high prediction confidence) do not imply low multiplicity. Our consistency measure provides better insight into the multiplicity of predictions compared to the predicted probabilities or drop-out prediction. Appendix D for visualizations on other Datasets.

Baselines: For comparison, we include the following baselines: *1) Prediction probability* $f(x)$ which measures the confidence of the model in predicting a given class. *2) Binary Drop-Out Method* (Hsu et al., 2024): Since there are no other baselines, we adapt this Drop-out method for TabLLMs. This method drops random weights of the model to explore models in the Rashomon set (i.e., set of competing models) without retraining several models. For a fair comparison, we compare our method (sampling k points in the neighborhood of our data point in the embedding space, and computing the consistency measure) to theirs (averaging the predictions of k models with different dropped-out weights). Note that these require the same number of inferences, hence complexity for both methods are around the same.[1] *Here are the experiments we conducted*:

• We plot the evaluated multiplicity against our consistency measure, predicted probabilities, and the drop-out method. See Figure 3 for illustration on the Adult 128 shot (BigScience T0 model). For the Bank, Diabetes, and German Credit dataset refer to Figure 4, 5, 6 in Appendix D.

• We compute the absolute spearman correlation between the consistency measures and various multiplicity evaluation metrics (128-shot setting on all datasets presented in Table 2). Full results on BigScience T0 model including 64 and 512 shot cases are presented in Table 5 in Appendix D. Results for Google FLAN-T5 model are presented in Table 6 in Appendix D.

Hyperparameter Selection and Ablations. Based on our theoretical results, choosing a larger *sample size* k is advantageous as it ensures the consistency guarantee holds with high probability. However, this also increases the computational cost of model inference. In our experiments, we set $k = 30$, the maximum number that fits into one inference pass on the GPU. For the *neighborhood radius* σ, we sampled perturbed points from a truncated Gaussian distribution with a variance of 0.01, which consistently performed well across all experiments. To guide the choice of σ, one could consider the spread of training samples. For the drop-out rate in the baseline, we use $p = 0.1$ following the recommendation in their paper (Hsu et al., 2024). The choice of δ in the competing set $\mathcal{F}_\delta$ is application-dependent; in our study, we used $\delta = 0.02$, corresponding to a 2% margin of accuracy deviation. Evaluating multiplicity by refining multiple models is computationally expensive. Thus, we limited our study to 40 models. To evaluate the impact of varying key parameters, *we conducted the following ablation studies*:

• We perform an ablation study on the sample size k, observing improved performance with increasing k. Detailed results are provided in Table 7 in Appendix D.

• We explore the effect of varying the perturbation radius σ. Results of this ablation study are summarized in Figure 7 and Table 8 in Appendix D. Best performance is observed at $\sigma = 10^{-2}$. When σ is too small (e.g., 10^{-4}), we basically sample (almost) the same points and our consistency measure is not more informative than the prediction probability. When σ is too large (e.g., 10^{-1}), one loses all information about the data point.

[1] Hsu et al. (2024) requires a prior check to ensure all dropped-out models (the models to be aggregated) are competing models (in terms of accuracy or loss), hence our method would be more computationally efficient under the same k.

Table 2: This table reports the Absolute Spearman Correlation between the consistency measure and various multiplicity evaluation metrics for 128 shots on the datasets. In most cases, our consistency measure $S_{k,\sigma}(x, f)$ shows a higher correlation with these multiplicity measures compared to predicted probabilities and drop-out method, indicating that the consistency measure $S_{k,\sigma}(x, f)$ better informs about the multiplicity than other measures. See full Table 5 with 64 and 512 shot cases in Appendix D.

Dataset	Number of Shots	Measure	Arbitrariness	Pairwise Disagreement	Prediction Variance	Prediction Range
Adult	128	Consistency	**0.80**	**0.96**	**0.84**	**0.91**
		Drop-Out	0.74	0.83	0.69	0.81
		Pred. Prob.	0.67	0.62	0.30	0.54
German	128	Consistency	0.54	0.54	**0.87**	**0.87**
		Drop-Out	0.50	0.56	0.74	0.84
		Pred. Prob.	**0.57**	**0.57**	0.86	0.86
Diabetes	128	Consistency	**0.92**	**0.95**	**0.93**	**0.95**
		Drop-Out	0.89	0.92	0.92	0.94
		Pred. Prob.	0.88	0.93	**0.93**	**0.95**
Bank	128	Consistency	**0.79**	**0.84**	**0.87**	**0.86**
		Drop-Out	0.62	0.70	0.75	0.51
		Pred. Prob.	0.54	0.57	0.73	0.62
Heart	128	Consistency	**0.89**	**0.90**	**0.97**	**0.87**
		Drop-Out	0.64	0.76	0.74	0.83
		Pred. Prob.	0.61	0.46	0.50	0.26
Car	128	Consistency	**0.97**	**0.91**	**0.93**	**0.94**
		Drop-Out	0.63	0.66	0.57	0.52
		Pred. Prob.	0.56	0.26	0.29	0.01

- We also evaluate the Drop-Out method with varying drop-out rates $p \in \{0.01, 0.1, 0.2, 0.5\}$. The correlation values between evaluated multiplicity and the consistency measures for the 512-shot setting on the Diabetes dataset are summarized in Table 9 in Appendix D. Our consistency measure outperforms the dropout method for all p values.

Discussions. Our multiplicity evaluation metrics, summarized in Table 1,3,4, reveal significant variability in model predictions across different fine-tuned variants, even when they exhibit similar accuracy. This multiplicity is not captured by merely examining predicted probabilities, as predictions with high confidence can still be susceptible to multiplicity (see Figure 3). Our consistency measure, $S_{k,\sigma}(x, f)$, was compared with prediction probabilities $f(x)$. The results, presented in Table 2,5,6, demonstrate that our consistency measure consistently shows mainly higher correlation with multiplicity metrics across all models and datasets compared to prediction probabilities and drop-out method. This indicates that $S_{k,\sigma}(x, f)$ is more informative than the baselines in informing the fine-tuning multiplicity. The drop-out method is however better than the prediction probabilities alone. We hypothesize that our method is more suitable for LLMs because the embedding space of LLMs is significantly smaller than the parameter space (possibly more informative also). The drop-out method might need significantly more inferences to compete due to this.

We study the unique nature of fine-tuning multiplicity in Tabular LLMs. Marx et al. (2020); Rudin et al. (2024) argue for the necessity of measuring and reporting multiplicity to better inform predictions. Traditional methods to measure multiplicity in classical ML are impractical for LLMs due to the computational challenge of retraining several fine-tuned models (Marx et al., 2020; Hsu & Calmon, 2022; Watson-Daniels et al., 2023). Our proposed measure, which requires only the given model and leverages the embedding space to inform multiplicity, addresses this issue. This approach reduces the complexity from retraining and inference to just inference, making it more feasible to apply in practice. Although, from our theoretical guarantee, a large k (number of sampled points) might be needed for accurate consistency estimation (particularly when β is large), it remains computationally more efficient than retraining multiple models. Our work provides practitioners with meaningful information about the multiplicity of predictions, which may lead them to carefully evaluate which predictions to trust and which to treat with caution. Our research has significant implications in several high-stakes applications, e.g., hiring, finance, education, etc., where inconsistent predictions can lead to distrust. A limitation of our work is that while we inform about fine-tuning multiplicity for a given sample, we do not resolve it. Future work could focus on developing methods to mitigate fine-tuning multiplicity, ensuring more consistent model predictions (see Appendix A for detailed discussion on Societal Impact and Limitations).

References

Spurthi Amba Hombaiah, Tao Chen, Mingyang Zhang, Michael Bendersky, and Marc Najork. Dynamic language models for continuously evolving content. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2514–2524, 2021.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Dimitris Bertsimas, Kimberly Villalobos Carballo, Yu Ma, Liangyuan Na, Léonard Boussioux, Cynthia Zeng, Luis R Soenksen, and Ignacio Fuentes. Tabtext: a systematic approach to aggregate knowledge across tabular data structures. *arXiv preprint arXiv:2206.10381*, 2022.

Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. Consistent counterfactuals for deep models. *arXiv preprint arXiv:2110.03109*, 2021.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 850–863, 2022.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

Leo Breiman. Statistical modeling: The two cultures. *Quality control and applied statistics*, 48(1): 81–82, 2003.

Vinay Chamola, Vikas Hassija, A Razia Sulthana, Debshishu Ghosh, Divyansh Dhingra, and Biplab Sikdar. A review of trustworthy and explainable artificial intelligence (xai). *IEEE Access*, 2023.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023a.

Zekai Chen, Mariann Micsinai Balan, and Kevin Brown. Language models are few-shot learners for prognostic prediction. *arXiv preprint arXiv:2302.12692*, 2023b.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2144–2155. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/coston21a.html.

Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1):26–43, 2022.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784, 2022.

Josip Djolonga, Frances Hubis, Matthias Minderer, Zachary Nado, Jeremy Nixon, Rob Romijnders, Dustin Tran, and Mario Lucic. Robustness Metrics, 2020. URL https://github.com/google-research/robustness_metrics.

Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. Robust counterfactual explanations for tree-based ensembles. In *International Conference on Machine Learning*, pp. 5742–5756. PMLR, 2022.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large language models (llms) on tabular data: Predic-tion, generation, and understanding-a survey. *arXiv preprint arXiv:2402.17944*, 2024.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Juan Felipe Gomez, Caio Vieira Machado, Lucas Monteiro Paes, and Flavio P Calmon. Algorithmic arbitrariness in content moderation. *arXiv preprint arXiv:2402.16979*, 2024.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *International Conference on Machine Learning*, pp. 12351–12367. PMLR, 2023.

Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Efficient estimation of the local robustness of machine learning models. *arXiv preprint arXiv:2307.13885*, 2023.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pp. 5549–5581. PMLR, 2023.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

Hsiang Hsu and Flavio Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems*, 35:28988–29000, 2022.

Hsiang Hsu, Ivan Brugere, Shubham Sharma, Freddy Lecue, and Chun-Fu Chen. Rashomongb: Analyzing the rashomon effect and mitigating predictive multiplicity in gradient boosting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hsiang Hsu, Guihong Li, Shaohan Hu, and Chun-Fu Chen. Dropout-based rashomon set exploration for efficient predictive multiplicity estimation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Sf2A2PUXO3.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Sukriti Jaitly, Tanay Shah, Ashish Shugani, and Razik Singh Grewal. Towards better serialization of tabular data for few-shot classification. *arXiv preprint arXiv:2312.12464*, 2023.

Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. *Advances in neural information processing systems*, 34:23928–23941, 2021.

Michael Kahn. Diabetes. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T59G.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571. PMLR, 2021.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*, 2024.

Xiangyang Li, Bo Chen, Lu Hou, and Ruiming Tang. Ctrl: Connect tabular and language model for ctr prediction. *arXiv preprint arXiv:2306.02841*, 2023.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep entity matching with pre-trained language models. *arXiv preprint arXiv:2004.00584*, 2020.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24 (253):1–15, 2023.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pp. 6765–6774. PMLR, 2020.

Kota Mata, Kentaro Kanamori, and Hiroki Arimura. Computing the collection of good models for rule lists. *arXiv preprint arXiv:2204.11285*, 2022.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Avanika Narayan, Ines Chami, Laurel Orr, Simran Arora, and Christopher Ré. Can foundation models wrangle your data? *arXiv preprint arXiv:2205.09911*, 2022.

Soma Onishi, Kenta Oono, and Kohei Hayashi. Tabret: Pre-training transformer-based tabular models for unseen columns. *arXiv preprint arXiv:2303.15747*, 2023.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pp. 809–818. PMLR, 2020.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Amazing things come from having many good models. *arXiv preprint arXiv:2407.04846*, 2024.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1827–1858, 2022.

Kacper Sokol, Meelis Kull, Jeffrey Chan, and Flora Dilys Salim. Fairness and ethics under model multiplicity in machine learning. *arXiv preprint arXiv:2203.07139*, 2022.

Kacper Sokol, Meelis Kull, Jeffrey Chan, and Flora Dilys Salim. Cross-model fairness: Empirical study of fairness and ethics under model multiplicity, 2023.

Karthik Sridharan. A gentle introduction to concentration inequalities. *Dept. Comput. Sci., Cornell Univ., Tech. Rep*, 2002.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 645–654, 2024.

Boris van Breugel and Mihaela van der Schaar. Why tabular foundation models should be a research priority. *arXiv preprint arXiv:2405.01147*, 2024.

Voigt. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.

Ruiyu Wang, Zifeng Wang, and Jimeng Sun. Unipredict: Large language models are universal tabular predictors. *arXiv preprint arXiv:2310.03266*, 2023.

Zifeng Wang, Chufan Gao, Cao Xiao, and Jimeng Sun. Meditab: Scaling medical tabular data predictors via data consolidation, enrichment, and refinement, 2024b.

Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification, 2023.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey, 2024.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems*, 35:14071–14084, 2022.

Jiahuan Yan, Bo Zheng, Hongxia Xu, Yiheng Zhu, Danny Chen, Jimeng Sun, Jian Wu, and Jintai Chen. Making pre-trained language models great on tabular prediction. *arXiv preprint arXiv:2403.01841*, 2024.

Yazheng Yang, Yuqi Wang, Sankalok Sen, Lei Li, and Qi Liu. Unleashing the potential of large language models for predictive tabular tasks in data science. *arXiv preprint arXiv:2403.20208*, 2024.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*, 2020.

Yuwei Yin, Yazheng Yang, Jian Yang, and Qi Liu. Finpt: Financial risk prediction with profile tuning on pretrained foundation models. *arXiv preprint arXiv:2308.00065*, 2023.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. *arXiv preprint arXiv:2310.17513*, 2023.

Han Zhang, Xumeng Wen, Shun Zheng, Wei Xu, and Jiang Bian. Towards foundation models for learning on tabular data. *arXiv preprint arXiv:2310.07338*, 2023.

A Social Impact and Limitations

Limitations. While our work provides a measure to assess fine-tuning multiplicity, it does not directly resolve this issue. Future research could focus on mitigation methods to ensure more consistent model predictions. A key constraint is the applicability to higher-dimensional datasets due to the limited context window size of current LLMs, though extending context windows is an active area of research (Peng et al., 2023; Chen et al., 2023a). Additionally, our method's performance can be sensitive to hyperparameters, such as sample size and neighborhood radius; incorrect choices may lead to an inaccurate assessment of robustness. Our approach also assumes access to the embedding space, limiting its application to open-source models. Furthermore, the bound in Theorem 1 is not directly computable. Estimating these unknowns such as β, ϵ' could be a direction for future work. Despite these limitations, our measure serves as a crucial step toward understanding and quantifying fine-tuning multiplicity, laying the groundwork for future advancements.

Broader Societal Impacts. The application of LLMs to tabular data, particularly in high-stakes domains such as finance and healthcare, presents both opportunities and risks (Bommasani et al., 2021). Our work aims to address one of the critical challenges associated with these models: the instability introduced when fine-tuning large models on small datasets. This instability, manifested as overfitting and multiplicity, can undermine the reliability of model predictions in scenarios where consistency is crucial. By measuring multiplicity, our work contributes to the responsible deployment of LLMs in domains where erroneous predictions can have severe consequences (Bommasani et al., 2021; Creel & Hellman, 2022).

Tabular data remains a dominant modality in many critical fields, yet it has received less research attention compared to text and image data (Hegselmann et al., 2023). Recent work van Breugel & van der Schaar (2024) argues that developing reliable foundation models for tabular data should be a research priority. Our deliberate focus on Tabular LLMs aligns with this perspective, as we aim to bridge a significant gap in the current research landscape. While it is understood that some degree of multiplicity is inherent in fine-tuned models, understanding its nature and impact is essential for building trust in Tabular LLMs.

Our approach also supports *regulatory compliance* by enhancing transparency and accountability in automated decision-making systems. Quantifying prediction robustness aligns with regulations such as the General Data Protection Regulation (GDPR) (Voigt, 2017) and upcoming AI legislation, which increasingly demand explainable and reliable AI models (Chamola et al., 2023). While LLMs are more computationally expensive than traditional models, our method reduces the costs of assessing multiplicity. By avoiding repeated retraining, it enhances *cost efficiency* and *minimizes environmental impact*, lowering both energy consumption and carbon footprint (Luccioni et al., 2023).

Furthermore, observing the nature of fine-tuning multiplicity in Tabular LLMs pave the way for future research into model stability. It also *facilitates continual learning* by informing the robustness of a prediction to potential model updates in a dynamic environments where data constantly evolves (Amba Hombaiah et al., 2021; Wu et al., 2024; Wang et al., 2024a). Lastly, our work could play a role in mitigating *fairwashing risks* and *explanation bias* (Black et al., 2022; Sokol et al., 2023; Rudin et al., 2024). This transparency is crucial for maintaining ethical standards and trustworthiness in AI deployment (Chamola et al., 2023).

B Additional Intuition Behind the Consistency Measure

The consistency measure quantifies the robustness of a model's prediction at a specific point x by examining the model's behavior in the local neighborhood of x within the embedding space. Our measure is motivated by our observations on synthetic data experiments where the model was exhibiting noisy and non-smooth patterns in the decision space.

Local Averaging: The term $\frac{1}{k}\sum_{x_i \in N_{x,k}} f(x_i)$ represents the average prediction of the model on points sampled from a neighborhood around x. This captures the general tendency of the model in the vicinity of x.

Variability Penalization: The term $\frac{1}{k}\sum_{x_i \in N_{x,k}} |f(x) - f(x_i)|$ computes the average absolute difference between the model's prediction at x and its predictions at neighboring points. Subtracting this from the local average penalizes the consistency score when there is high variability in a neighborhood in spite of high local mean, reflecting instability in the model's predictions around x.

By combining these two terms, $S_{k,\sigma}(x, f)$ provides a measure that is high when the model's predictions are both strong (i.e., high average prediction) and stable (i.e., low variability) in the neighborhood of x. The metric is designed to capture the local stability of the model's predictions, which is critical in assessing robustness to fine-tuning multiplicity.

Consistency Interpretation of $f(x)$ and $2f(x_i) - f(x)$: This interesting structure of our consistency measure is not a heuristic design but arises directly from the reverse triangle inequality step in the proof of our theoretical consistency guarantee (Theorem 1): $|f(x)| \geq |f(x_i)| - |f(x_i) - f(x)|$.

When $f(x_i) \geq f(x)$, the contribution to the consistency score is $f(x)$, indicating that the neighborhood prediction $f(x_i)$ reinforces and supports the robustness of prediction $f(x)$.

When $f(x_i) < f(x)$, the contribution becomes $2f(x_i) - f(x)$. If $f(x_i)$ is significantly less than $f(x)$, the term $2f(x_i) - f(x)$ becomes negative, penalizing the consistency score due to large discrepancies between $f(x)$ and its neighbor. However, if $f(x_i)$ is only slightly less than $f(x)$ (i.e., $f(x_i) > \frac{f(x)}{2}$), the term $2f(x_i) - f(x)$ remains positive, thereby contributing positively to the consistency measure. The intuition is that we only penalize significant drops in neighboring predictions and allow neighbors that closely support $f(x)$ prediction.

How consistency differs from existing robustness: Our focus on model multiplicity distinguishes this work from traditional robustness measures, which address different aspects of model behavior such as out-of-distribution (OOD) generalization, stability under natural perturbations, and uncertainty estimation (Djolonga et al., 2020). OOD generalization typically evaluates how well a model performs on data that differs from the training distribution (e.g., classifying objects seen from novel viewpoints or in cluttered settings). This is often quantified using test datasets with altered conditions or domain shifts, and methods like domain adaptation are employed to enhance robustness. Stability under natural perturbations assesses the sensitivity of predictions and predicted probabilities to small, random changes in the input, such as Gaussian noise or image transformations. Uncertainty estimation, on the other hand, focuses on calibrating the predicted probabilities to reflect true likelihoods, often using measures like Expected Calibration Error or entropy-based metrics to evaluate how well the model quantifies confidence in its predictions. While these methods provide valuable insights into different facets of robustness, their goals differ significantly from ours.

Han et al. (2023) is more closely related to our approach, as it quantifies robustness by measuring the fraction of consistent predictions within a local neighborhood. While both approaches leverage the neighborhood around a data point, the objectives diverge: Han et al. (2023) focuses on quantifying the probability of consistent predictions against perturbations to evaluate robustness to noise. In contrast, our measure captures the consistency of predictions (multiplicity) among competing models within the Rashomon set.

Additionally, our consistency measure's unique mean-variance nature further distinguishes it. Unlike existing metrics, it not only accounts for the average prediction within a neighborhood but also penalizes the variability in predictions. Moreover, we provide theoretical guarantees on the robustness of predictions with high consistency scores over a broad range of equally-well performing models.

C Proof of Theoretical Guarantee

Theorem 1 (Probabilistic Guarantee on Consistency)**.** *Given a data point x, a random model F' and consistency measure $S_{k,\sigma}(x, F')$. Then under Assumption 1, and $|\mathbb{E}[Z_i|F' = f', F = f]| \leq \epsilon'$, a prediction over a broad class of fine-tuned models satisfies:*

$$\Pr(F(x) \geq S_{k,\sigma}(x, F') - \epsilon) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8(\beta + \frac{2}{3}\epsilon)}\right), \tag{2}$$

for all $\epsilon > 2\epsilon'$, The probability is over the stochastic models F and F', and the random perturbations X_i's are random points sampled independently from a distribution over a hypersphere $B(x, \sigma)$.

Proof. To prove Theorem 1, we begin with Lemma 1.

Assume the fine-tuned models F belong to a discrete class of random variables. A specific model realization is represented as f_i for $i = 1, 2, \ldots, |\mathcal{F}_\delta|$, with the complete set denoted by $\mathcal{F} = \{f_1, f_2, \ldots, f_{|\mathcal{F}|}\}$. Each model f_i is selected with probability p_i, where $\sum_{i=1}^{|\mathcal{F}_\delta|} p_i = 1$.

Lemma 1 (Neighborhood Divergence Bound)**.** *Given the neighborhood discrepancy Z, under Assumption 1, for any $\tilde{\epsilon} > \epsilon' > 0$, we have:*

$$\Pr(Z \geq \epsilon' + \tilde{\epsilon}) \leq \exp\left(\frac{-k(\tilde{\epsilon} + \epsilon')^2}{8\beta + \frac{16}{3}(\tilde{\epsilon} + \epsilon')}\right). \tag{3}$$

Let $Z = \frac{1}{k}\sum_{i=1}^{k} Z_i$. We show that $\mathbb{E}\left[Z\right] = 0$:

$$\mathbb{E}[Z] \overset{(a)}{=} \mathbb{E}_{X_i}\left[\mathbb{E}_{F|X_i}\left[\frac{1}{k}\sum_{i=1}^{k}(F'(X_i) - F(X_i) - |F(X_i) - F(x)| + |F'(X_i) - F'(x)|)\right]\right] \tag{4}$$

$$\overset{(b)}{=} \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{X_i}[\mathbb{E}_{F|X_i}[(F'(X_i) - F(X_i) - |F(X_i) - F(x)| + |F'(X_i) - F'(x)|)]] \tag{5}$$

$$\overset{(c)}{=} \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{X_i}\left[\mathbb{E}[F'(X_i)|X_i] - \mathbb{E}[F(X_i)|X_i] - \mathbb{E}[|F(X_i) - F(x)||X_i]\right. \tag{6}$$

$$\left.+\,\mathbb{E}[|F'(X_i) - F'(x)||X_i]\right] \tag{7}$$

$$\overset{(d)}{=} \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{X_i}\left[\mathbb{E}[F(X_i)|X_i] - \mathbb{E}[F(X_i)|X_i] - \mathbb{E}[|F(X_i) - F(x)||X_i]\right. \tag{8}$$

$$\left.+\,\mathbb{E}[|F(X_i) - F(x)||X_i]\right] = 0 \tag{9}$$

Here *(a)* holds from applying the law of total expectation. *(b)* Distributing the expectation over the summation. *(c)* Applying the linearity of expectations inside the inner expectation. *(d)* From Assumption 1, F and F' are identically distributed given X_i, hence $\mathbb{E}[F'(X_i)|X_i] = \mathbb{E}[F(X_i)|X_i]$ and $\mathbb{E}[|F'(X_i) - F'(x)||X_i] = \mathbb{E}[|F(X_i) - F(x)||X_i]$. The terms cancel each other out, resulting in $\mathbb{E}[Z] = 0$. The rest of the proof leverages Bernstien's Inequality:

Lemma 2 (Bernstein Inequality)**.** *For a given random variable X_i such that $\Pr(|X_i| \leq c) = 1$, and $\beta = \frac{1}{k}\sum_{i=1}^{k}\mathrm{Var}\left[X_i\right]$ then, for any $\varepsilon > 0$,*

$$\Pr\left(\left|\frac{1}{k}\sum_{i=1}^{k}X_i - \mathbb{E}(X_i)\right| > \varepsilon\right) \leq 2\exp\left(\frac{-k\varepsilon^2}{2\beta + \frac{2c\varepsilon}{3}}\right). \tag{10}$$

See Sridharan (2002) for detailed proof of Bernstein's Inequality.

Observe that $|Z_i| = |F'(X_i) - F(X_i) - |F(X_i) - F(x)| + |F'(X_i) - F'(x)|| \leq 2$. Hence, we have:

$$\Pr\left(|Z - \mathbb{E}[Z|F' = f', F = f]| \geq \tilde{\epsilon} \mid F' = f', F = f\right) \leq 2\exp\left(-\frac{k\tilde{\epsilon}^2}{2\beta + \frac{4}{3}\tilde{\epsilon}}\right)$$

where $\frac{1}{k}\sum_{i=1}^{k}\mathrm{Var}\left[Z_i|F' = f', F = f\right] \leq \beta$ from Assumption 1.

Given $|\mathbb{E}\left[Z|F' = f', F = f\right] - \mathbb{E}\left[Z\right]| < \epsilon'$ and $\mathbb{E}\left[Z\right] = 0$,

we have $-\epsilon' < \mathbb{E}\left[Z|F' = f', F = f\right] < \epsilon'\ \forall f, f'$. Now observe that:

$$\Pr(Z \geq \epsilon' + \tilde{\epsilon}|F' = f', F = f) \overset{\text{(a)}}{\leq} \Pr(Z \geq \mathbb{E}\left[Z|F' = f', F = f\right] + \tilde{\epsilon}|F' = f', F = f) \tag{11}$$

$$\leq \exp\left(\frac{-k\tilde{\epsilon}^2}{2\beta + \frac{4}{3}\tilde{\epsilon}}\right). \tag{12}$$

Here, (a) holds since $\mathbb{E}\left[Z|F' = f', F = f\right] < \epsilon'$. The event on the left is a subset of that on the right. Therefore, the probability of the event $\{Z \geq \epsilon' + \tilde{\epsilon}\}$ occurring cannot be more than the probability of the event $\{Z \geq \mathbb{E}[Z|F' = f', F = f] + \tilde{\epsilon}\}$ occurring.

$$\begin{aligned}
\Pr(Z \geq \epsilon' + \tilde{\epsilon}) &\overset{\text{(b)}}{=} \sum_{i,j} \Pr(Z \geq \epsilon' + \tilde{\epsilon} | F' = f_i, F = f_j) \Pr(F' = f_i, F = f_j) && (13)\\
&\overset{\text{(c)}}{\leq} \exp\left(\frac{-k\tilde{\epsilon}^2}{2\beta + \frac{4}{3}\tilde{\epsilon}}\right) \sum_{i,j} \Pr(F' = f_i, F = f_j) && (14)\\
&= \exp\left(\frac{-k\tilde{\epsilon}^2}{2\beta + \frac{4}{3}\tilde{\epsilon}}\right) && (15)\\
&\overset{\text{(d)}}{\leq} \exp\left(\frac{-k(\tilde{\epsilon} + \epsilon')^2}{8\beta + \frac{16}{3}(\tilde{\epsilon} + \epsilon')}\right) && (16)
\end{aligned}$$

Here, (b) holds from the law of total probability. Next, (c) follows from Equation 12. Finally, (d) holds from using the inequality $4\tilde{\epsilon}^2 > (\tilde{\epsilon} + \epsilon')^2$ which holds for $\tilde{\epsilon} > \epsilon' > 0$ at the numerator and $\tilde{\epsilon} \leq \tilde{\epsilon} + \epsilon'$ at the denominator. Setting $\epsilon = \tilde{\epsilon} + \epsilon'$.

We have:

$$\Pr\left(\frac{1}{k}\sum_{i=1}^{k} F(X_i) \geq \frac{1}{k}\sum_{i=1}^{k}\big(F'(X_i) - |F'(X_i) - F'(x)| + |F(X_i) - F(x)|\big) - \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8\beta + \frac{16}{3}\epsilon}\right). \tag{17}$$

Observe that $F(x) \geq F(x_i) - |F(x_i) - F(x)|$. This applies directly from the reverse triangle inequality, i.e., for any real numbers a and b, we have:$|a| \geq |b| - |a - b|$.

Hence,

$$F(x) \geq \frac{1}{k}\sum_{i=1}^{k}(F(X_i) - |F(X_i) - F(x)|) \tag{18}$$

Therefore, plugging equation 18 into equation 17, we have:

$$\Pr\big(F(x) \geq \frac{1}{k}\sum_{i=1}^{k}(F'(X_i) - |F'(X_i) - F'(x)| + |F(X_i) - F(x)| - |F(X_i) - F(x)| - \epsilon)\big) \tag{19}$$

$$= \Pr\big(F(x) \geq \frac{1}{k}\sum_{i=1}^{k}(F'(X_i) - |F'(X_i) - F'(x)|) - \epsilon\big) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8\beta + \frac{16}{3}\epsilon}\right). \tag{20}$$

Given $S_{k,\sigma}(x, F') = \frac{1}{k}\sum_{i=1}^{k}(F(X_i) - |F'(x) - F'(X_i)|)$, we have:

$$\Pr\big(F(x) \geq S_{k,\sigma}(x, F') - \epsilon\big) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8\beta + \frac{16}{3}\epsilon}\right). \tag{21}$$

Remark 2 (Randomness of F and F')**.** *In Theorem 1, we consider both F and F' as random to capture the variability inherent in the fine-tuning process. This approach models the real-world scenario where different fine-tuning runs can lead to different models due to changes in random seeds or training conditions. Fixing $F' = f'$ will require an alternate assumption instead of Assumption 1 that the predictions of other models are centered around f', i.e., $\mathbb{E}[F(X)|X = x] = f'(x)$. While alternative bounds using a fixed F' could provide valuable insights, our current approach aims to capture the randomness of the initial fine-tuned model (F') and understand robustness across the broader distribution of possible fine-tuned models.*

Remark 3 (Variance of Z)**.** *The variance of Z is bounded for functions F and F', and we could have used the trivial bound of $\beta \leq 2$ in our guarantee. However, we anticipate that β is significantly smaller, particularly on the data manifold, because F and F' are models fine-tuned from the same pretrained on the same dataset. For samples lying on the data manifold—where realistic samples exist—we expect several models (from the same pretrained model) fine tuned on the same dataset with different training seed to exhibit "similar" prediction probabilities. However, fine-tuned models* *can differ significantly in regions outside the data manifold, as the absence of training samples in these areas means there is no shared information to constrain their behavior.*

The upper bound of $\beta \leq 2$ can be used to determine the worst-case sample size k needed to ensure the guarantees to some certifiable gap. This provides a conservative estimate that remains applicable even in the absence of precise parameter knowledge.

□

D EXPANDED EXPERIMENTAL RESULTS

D.1 RELEVANT DEFINITION

Definition 6 (Spearman Correlation)**.** *Spearman's correlation, Spearman(X, Y), measures the strength and direction of a monotonic relationship between two variables. It is the Pearson correlation coefficient of their ranked values.*

Given n pairs (X_i, Y_i), it is computed as:

$$\textit{Spearman}(X,Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = \frac{\textit{cov}(\textit{rank}(X), \textit{rank}(Y))}{\sigma_{\textit{rank}(X)}\sigma_{\textit{rank}(Y)}},$$

where d_i is the difference between the ranks of X_i and Y_i. The value ranges from -1 (perfect negative monotonicity) to 1 (perfect positive monotonicity), with 0 indicating no monotonic relationship.

D.2 DATASET DETAILS

Adult Dataset. The Adult dataset (Becker & Kohavi, 1996), also known as the "Census Income" dataset, is used for predicting whether an individual earns more than \$50,000 annually based on various demographic attributes. It consists of 48,842 instances with 14 attributes, including age, work class, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, and native country. The dataset is commonly used in classification tasks.

German Credit Dataset. The German Credit dataset (Hofmann, 1994) is used for credit risk evaluation. It consists of $1,000$ instances with 20 attributes, which include personal information, credit history, and loan attributes. The target variable indicates whether the credit is good or bad. This dataset is often used for binary classification problems and helps in understanding the factors affecting creditworthiness. The dataset is commonly used in classification tasks.

Diabetes Dataset. The Diabetes dataset Kahn is used for predicting the onset of diabetes based on diagnostic measurements. It contains 768 instances with 8 attributes, including the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The target variable indicates whether the individual has diabetes. The dataset is commonly used in classification tasks.

Bank Dataset. The Bank dataset (Moro et al., 2014) is used for predicting whether a client will subscribe to a term deposit based on data from direct marketing campaigns of a Portuguese bank. It includes 45,211 instances in the training set and 18 attributes, such as age, job type, marital status, education, credit balance, housing loan status, and contact details from the marketing campaigns. The target variable indicates whether the client subscribed to the term deposit. This dataset is commonly used in binary classification tasks.

Heart Dataset. The Heart dataset contains data from four different hospitals. It includes 918 patients, each represented by 11 clinical variables, with the task being a binary classification of coronary artery disease. Among the patients, 508 are labeled positive for the condition.

Car Dataset. The Car dataset contains entries describing various cars characterized by six attributes. The task is a classification problem aimed at evaluating the state of each car. The dataset comprises 1,728 examples.

Table 3: Multiplicity Evaluation Metrics for Different Datasets and Number of Shots. Evaluated on 40 fine-tuned **BigScience T0** models on **LORA** using different random seeds. Multiplicity observed in predictions across different fine-tuned model, even when models exhibit similar accuracy (in this setting $\delta = 0.02$).

Dataset	**No. Shots**	**Multiplicity Evaluation Metrics (BigScience T0)**					
		Arbitrariness	**Discrepancy**	**Avg. Pairwise Disagreement**	**Avg. Pred. Variance**	**Avg. Pred. Range**	**Avg. Model Accuracy**
Adult	64	11%	6%	9%	0.01	0.11	83%
	128	10%	9%	6%	0.01	0.10	84%
	512	11%	3%	10%	0.01	0.12	85%
German	64	19%	10%	6%	0.04	0.40	70%
	128	17%	11%	6%	0.01	0.16	71%
	512	21%	14%	8%	0.03	0.26	72%
Diabetes	64	20%	13%	11%	0.04	0.21	70%
	128	16%	14%	11%	0.08	0.14	73%
	512	19%	13%	11%	0.04	0.17	76%
Bank	64	13%	9%	7%	0.01	0.28	66%
	128	14%	9%	7%	0.03	0.21	73%
	512	14%	8%	7%	0.03	0.22	78%

Table 4: Evaluated Multiplicity for Different Datasets and Number of Shots. Evaluated on 40 fine-tuned **FLAN-T5** models using **Tfew** recipe with different random seeds. Multiplicity observed in predictions across different fine-tuned models, even when models exhibit similar accuracy (in this setting $\delta = 0.02$). The accuracy of FLAN T5 model on the dataset is less than the BigScience T0 model observed in Table 1.

Dataset	**No. Shots**	**Multiplicity Evaluation Metrics (Google FLAN-T5)**					
		Arbitrariness	**Discrepancy**	**Avg. Pairwise Disagreement**	**Avg. Pred. Variance**	**Avg. Pred. Range**	**Avg. Model Accuracy**
Adult	64	13.96%	6.93%	5.05%	0.010	0.139	74.25%
	128	8.81%	3.84%	3.39%	0.008	0.091	77.50%
	512	12.02%	5.71%	4.49%	0.012	0.123	79.17%
German	64	18.50%	11.00%	6.19%	0.015	0.194	64.85%
	128	30.00%	13.50%	10.47%	0.031	0.287	69.25%
	512	35.50%	16.50%	12.88%	0.041	0.362	69.40%
Diabetes	64	15.58%	7.79%	6.23%	0.016	0.170	68.18%
	128	11.69%	5.84%	4.81%	0.012	0.129	59.29%
	512	21.43%	9.74%	7.37%	0.022	0.207	69.55%
Bank	64	12.86%	7.46%	4.69%	0.003	0.125	66.96%
	128	17.95%	6.90%	6.59%	0.006	0.165	65.94%
	512	17.17%	6.61%	6.24%	0.017	0.173	79.40%

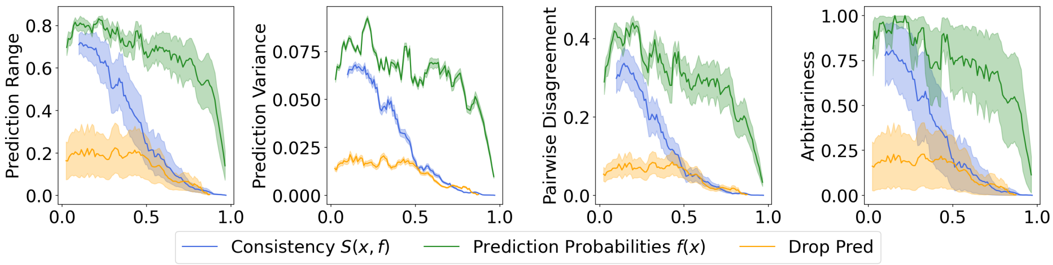


Figure 4: Evaluated multiplicity (assessed on 40 retrained models) versus our consistency measure (evaluated on one model) for the **512-shot** setting on the **Bank dataset**. The plots demonstrate that high consistency values correspond to low multiplicity across various multiplicity evaluation metrics. Predictive probabilities and Drop-Out not providing any providing any useful insight into multiplicity.

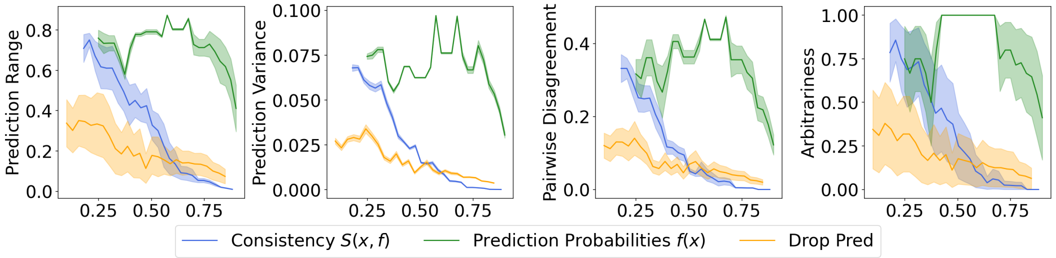


Figure 5: Evaluated multiplicity (assessed on 40 retrained models) versus our consistency measure (evaluated on one model) for the **512-shot** setting on the **Diabetes dataset**. The plots demonstrate that high consistency values correspond to low multiplicity across various multiplicity evaluation metrics. Predictive probabilities not providing any providing any useful insight about multiplicity. The drop-out method performs better than predictive probabilities but still worse than consistency.

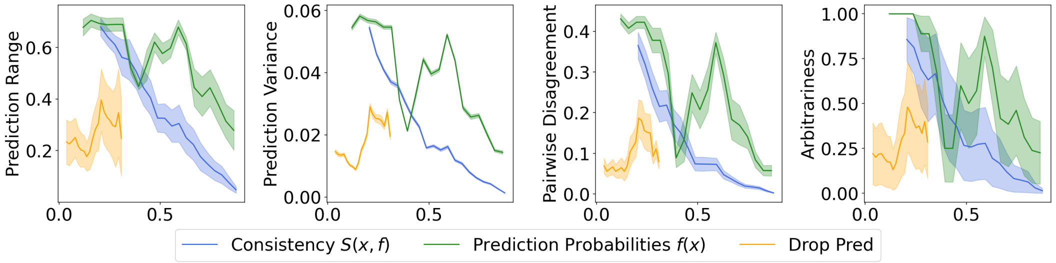


Figure 6: Evaluated multiplicity (assessed on 40 retrained models) versus our consistency measure (evaluated on one model) for the **512-shot** setting on the **German Credit dataset**. The plots demonstrate that high consistency values correspond to low multiplicity across various multiplicity evaluation metrics. In this setting Prediction probability is performing competitively. But generally consistency measure provides better insight into the multiplicity of predictions compared to the predicted probabilities. The drop-out method is performing significantly worse than the other two measures.

D.3 Experimental Setup

Our experiments were conducted using the BigScience T0 and Google Flan T5 models fine-tuned on four datasets: German Credit, Diabetes, Bank, and Adult Income. We explored the performance and robustness of the fine-tuned models in few-shot scenarios. The number of shots was set to 64,128, and 512 for each dataset. To evaluate model multiplicity and consistency, we fine-tuned 40 models with different random seeds for each dataset and recorded their predictions. The training process involved setting the batch size to 2 for smaller training sizes and 8 for larger sizes. The learning rate was set to 0.003. For each dataset, we determined the number of training steps adaptively based on the number of shots, ensuring sufficient iterations for model convergence. Specifically, for the number of shots-shot settings, the training steps were calculated as 20 × (number of shots/batch size). All experiments were performed on 2 NVIDIA RTX A4500 and 4 NVIDIA RTX 6000 GPUs. To ensure reproducibility and robustness of the results, different random seeds (i.e., 2, 4, 8, etc) were used for each fine-tuning iteration. For fine-tuning with LORA we use a rank of 4.

Remark 4. *Given the infeasibility of computing the exact size of $|\mathcal{F}_\delta|$ due to its potentially vast model space, we employ an expensive sampling approach, i.e., fine-tuning with various seeds. We select a finite number of models from $\mathcal{F}_\delta$ for practical evaluation, allowing us to evaluate the multiplicity metrics. It is very computationally expensive to fine-tune several models to evaluate multiplicity. This motivates the need for a measure to quantify consistency given one model.*

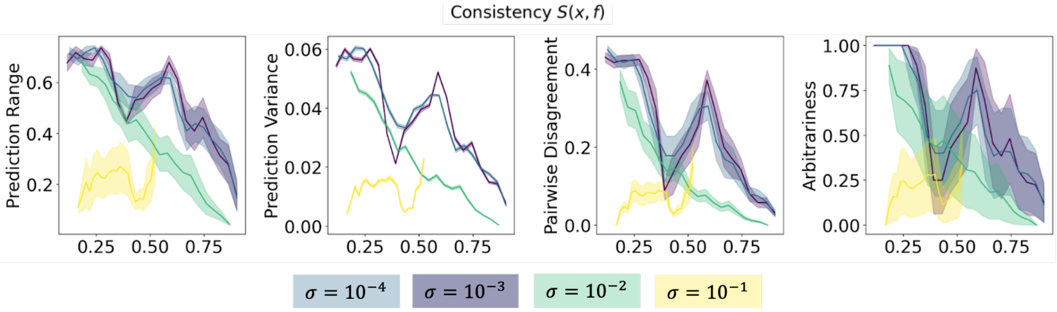


Figure 7: **Ablation study on different σ values**: The chosen value of $\sigma = 0.01$ yields the best performance across all evaluation metrics. Smaller values of σ (e.g., $\sigma = 10^{-4}$) result in perturbations that are too close to the original data points, leading to similar outcomes as prediction probability alone, as the sampled points are nearly identical. On the other hand, larger values (e.g., $\sigma = 10^{-2}$) produce overly noisy perturbations, rendering the results uninformative.

Table 5: This table reports the Spearman correlation between the consistency measure, predicted probabilities, and the drop-out method with various multiplicity evaluation metrics for different numbers of shots on the Adult, German Credit, Diabetes, Heart, Car, and Bank datasets (**BigScience T0 fine-tuned using Tfew recipe**). In most cases, the consistency measure $S_{k,\sigma}(x, f)$ shows a higher correlation with these multiplicity measures compared to predicted probabilities and drop-out, indicating that the consistency measure $S_{k,\sigma}(x, f)$ better informs about the multiplicity than the other measures do. The dropout method performing better than naive predicted probability.

Dataset	Number of Shots	Measure	Arbitrariness	Pairwise Disagreement	Prediction Variance	Prediction Range
Adult	64	Consistency	**0.95**	**0.90**	**0.91**	**0.89**
		Drop-Out	0.83	0.78	0.81	0.87
		Pred. Prob.	0.67	0.66	0.50	0.62
	128	Consistency	**0.80**	**0.96**	**0.84**	**0.91**
		Drop-Out	0.74	0.83	0.69	0.81
		Pred. Prob.	0.67	0.62	0.30	0.54
	512	Consistency	**0.90**	**0.86**	**0.93**	**0.92**
		Drop-Out	0.78	0.78	0.88	0.88
		Pred. Prob.	0.70	0.69	0.56	0.72
German Credit	64	Consistency	0.95	0.95	**0.98**	**0.84**
		Drop-Out	0.73	0.71	0.82	0.76
		Pred. Prob.	**0.99**	**0.99**	0.80	0.79
	128	Consistency	0.54	0.54	**0.87**	**0.87**
		Drop-Out	0.50	0.56	0.74	0.84
		Pred. Prob.	**0.57**	**0.57**	0.86	0.86
	512	Consistency	0.59	0.60	**0.87**	**0.86**
		Drop-Out	**0.69**	**0.67**	0.72	0.65
		Pred. Prob.	0.54	0.56	0.83	0.82
Diabetes	64	Consistency	**0.45**	**0.51**	0.31	0.23
		Drop-Out	0.30	0.19	**0.54**	**0.46**
		Pred. Prob.	0.03	0.38	0.04	0.08
	128	Consistency	**0.92**	**0.95**	**0.93**	**0.95**
		Drop-Out	0.89	0.92	0.92	0.94
		Pred. Prob.	0.88	0.93	**0.93**	**0.95**
	512	Consistency	**0.80**	**0.89**	0.74	0.68
		Drop-Out	0.74	0.83	**0.75**	**0.74**
		Pred. Prob.	0.21	0.23	0.24	0.30
Bank	64	Consistency	**0.83**	**0.78**	**0.81**	**0.80**
		Drop-Out	0.79	0.77	0.77	**0.80**
		Pred. Prob.	0.70	0.69	0.56	0.74
	128	Consistency	**0.79**	**0.84**	**0.87**	**0.86**
		Drop-Out	0.62	0.70	0.75	0.51
		Pred. Prob.	0.54	0.57	0.73	0.62
	512	Consistency	**0.91**	**0.92**	**0.91**	**0.87**
		Drop-Out	0.90	0.89	0.87	0.84
		Pred. Prob.	0.71	0.68	0.81	0.76
Heart	64	Consistency	0.98	0.86	0.98	0.98
		Drop-Out	0.56	0.48	0.54	0.56
		Pred. Prob.	0.70	0.21	0.30	0.69
	128	Consistency	**0.89**	**0.90**	**0.97**	**0.87**
		Drop-Out	0.64	0.76	0.74	0.83
		Pred. Prob.	0.61	0.46	0.50	0.26
	512	Consistency	0.89	0.95	0.86	0.95
		Drop-Out	0.94	0.90	0.90	0.94
		Pred. Prob.	0.80	0.65	0.48	0.35
Car	64	Consistency	0.76	0.69	0.86	0.75
		Drop-Out	0.85	0.83	0.96	0.97
		Pred. Prob.	0.83	0.83	0.40	0.83
	128	Consistency	0.97	0.91	0.93	0.94
		Drop-Out	0.63	0.66	0.57	0.52
		Pred. Prob.	0.56	0.26	0.29	0.01
	512	Consistency	.68	0.59	0.56	0.67
		Drop-Out	0.98	0.96	0.95	0.93
		Pred. Prob.	0.91	0.94	0.72	0.86

Table 6: This table reports the Spearman correlation between the predicted probabilities, drop-out method, and the consistency measure with various multiplicity evaluation metrics for different numbers of shots on the Adult, German Credit, Diabetes, and Bank datasets (**Flan T5 model fine-tuned using Tfew recipe**). In most cases, the consistency measure shows a higher correlation with these multiplicity measures compared to predicted probabilities and drop-out, indicating that the consistency measure better informs about the multiplicity than the other measures do. The dropout method performs competitively in some cases.

Dataset	Number of Shots	Measure	Arbitrariness	Pairwise Disagreement	Prediction Variance	Prediction Range
Adult	64	Pred. Prob.	0.62	0.67	**0.72**	0.56
		Drop-Out	0.60	0.65	0.67	0.57
		Consistency	**0.63**	**0.72**	**0.72**	**0.60**
	128	Pred. Prob.	0.75	0.74	0.65	0.75
		Drop-Out	0.85	0.78	0.83	0.75
		Consistency	**0.88**	**0.90**	**0.84**	**0.79**
	512	Pred. Prob.	0.78	0.68	0.42	0.45
		Drop-Out	0.78	0.78	0.42	0.45
		Consistency	**0.79**	**0.71**	**0.78**	**0.68**
German Credit	64	Pred. Prob.	0.27	0.04	0.27	0.17
		Drop-Out	0.73	0.45	0.60	0.17
		Consistency	**0.77**	**0.67**	**0.78**	**0.76**
	128	Pred. Prob.	0.85	0.76	0.85	0.91
		Drop-Out	0.86	0.91	0.85	0.91
		Consistency	**0.89**	**0.91**	**0.89**	**0.92**
	512	Pred. Prob.	0.42	0.29	0.27	0.19
		Drop-Out	0.43	0.36	0.28	0.33
		Consistency	**0.61**	**0.60**	**0.67**	**0.69**
Diabetes	64	Pred. Prob.	0.09	0.04	0.27	0.23
		Drop-Out	0.24	0.41	**0.54**	**0.50**
		Consistency	**0.27**	**0.55**	0.31	0.25
	128	Pred. Prob.	0.16	0.06	0.17	0.16
		Drop-Out	0.46	0.55	**0.54**	**0.63**
		Consistency	**0.52**	**0.57**	0.44	0.52
	512	Pred. Prob.	0.61	0.35	0.12	0.19
		Drop-Out	0.71	0.42	**0.42**	**0.51**
		Consistency	**0.79**	**0.40**	0.39	0.40
Bank	64	Pred. Prob.	0.26	0.04	0.27	0.17
		Drop-Out	0.24	0.60	0.60	**0.60**
		Consistency	**0.77**	**0.67**	**0.78**	0.76
	128	Pred. Prob.	0.45	0.54	0.73	0.62
		Drop-Out	0.62	0.70	0.75	**0.82**
		Consistency	**0.89**	**0.71**	**0.78**	0.84
	512	Pred. Prob.	0.42	0.29	0.27	0.11
		Drop-Out	0.44	0.29	**0.37**	**0.43**
		Consistency	**0.61**	**0.60**	0.30	0.38

Table 7: **Ablation study on different k values**: Correlation between our consistency measure (evaluated on a single model) and various measures of multiplicity for different sample sizes k on the Diabetes dataset (T0 model). We observe better performance with increasing k as suggested by our theoretical results. Larger sample size k values are advantageous, as they ensure that the guarantees hold with high probability. However, computational cost of model inference (forward pass) increases.

k	Prediction Range	Prediction Variance	Pairwise Disagreement	Arbitrariness
2	0.77	0.77	0.53	0.52
5	0.82	0.83	0.56	0.55
10	0.87	0.87	0.62	0.61
20	**0.89**	**0.88**	**0.70**	**0.79**

Table 8: **Ablation study on different σ values**: Correlation between our consistency measure (evaluated on one model) and various evaluation measures for different values of σ and evaluated multiplicity for Diabetes dataset and 128-shot case (T0 model). Best performance observed when $\sigma = 10^{-2}$. To guide the choice of σ, one could consider the spread of training data points in the embedding space (e.g., we use a value equivalent to 10% of the variance of the training data). For all our experiments, we used a fixed value of 0.01, which consistently worked well across different datasets and experiments. When σ is too small, we basically sample (almost) the same points and our consistency measure is not more informative than the prediction probability. When σ is too large, one loses all information about the data point.

σ	Prediction Range	Prediction Variance	Pairwise Disagreement	Arbitrariness
10^{-4}	0.82	0.83	0.84	0.80
10^{-3}	0.91	0.92	0.90	0.86
10^{-2}	**0.95**	**0.93**	**0.95**	**0.92**
10^{-1}	0.10	0.08	0.33	0.23

Table 9: This table reports the correlation between the consistency measure and various evaluated multiplicity for the 512-shot setting on the Diabetes dataset. The consistency measure $S_{k,\sigma}(x, f)$ shows a higher correlation with multiplicity compared to predicted probabilities and drop-out and ensemble method (Hsu et al., 2024), indicating that the consistency measure $S_{k,\sigma}(x, f)$ better informs about the multiplicity than the other measures.

Method	Arbitrariness	Pairwise Disagreement	Prediction Variance	Prediction Range
drop-out $p = 0.01$	0.21	0.23	0.27	0.28
drop-out $p = 0.1$	0.62	0.61	0.59	0.64
drop-out $p = 0.2$	0.74	0.36	0.53	0.54
drop-out $p = 0.5$	0.16	0.17	0.18	0.16
Pred Prob	0.21	0.23	0.24	0.30
Consistency (ours)	**0.80**	**0.89**	**0.74**	**0.68**